

On the Mathematical Relationship Between Layer Normalization and Dynamic Activation Functions

Felix Stollenwerk

AI Sweden

felix.stollenwerk@ai.se

Abstract

Layer normalization (LN) is an essential component of modern neural networks. While many alternative techniques have been proposed, none of them have succeeded in replacing LN so far. The latest suggestion in this line of research is a dynamic activation function called Dynamic Tanh (DyT). Although it is empirically well-motivated and appealing from a practical point of view, it lacks a theoretical foundation. In this work, we shed light on the mathematical relationship between LN and dynamic activation functions. In particular, we derive DyT from the LN variant RMSNorm, and show that a well-defined decoupling in derivative space as well as an approximation are needed to do so. By applying the same decoupling procedure directly in function space, we are able to omit the approximation and obtain the exact element-wise counterpart of RMSNorm, which we call Dynamic Inverse Square Root Unit (DyISRU). We demonstrate numerically that DyISRU reproduces the normalization effect on outliers more accurately than DyT does.

1 Introduction

Most modern neural network architectures contain normalization layers. These have been shown to have beneficial effects on model training, such as faster, more stable convergence and better results (see e.g. [Huang et al. \(2023\)](#)). The most widely used normalization layers nowadays, especially in transformers, are layer normalization ([Ba et al., 2016](#)) and RMSNorm ([Zhang and Sennrich, 2019](#)). Both employ activation statistics across the channels (or model dimension) of the neural network. More concretely, consider a hidden representation $x \in \mathbb{R}^C$, corresponding for instance to a single token in the case of an LLM. Layer normalization (LN) acts on x by centering and scaling,

$$y = \frac{x - \mu}{\sqrt{\sigma^2}} \quad (1)$$

using the (sample) mean and variance,

$$\mu = \frac{1}{C} \sum_{k=1}^C x_k, \quad \sigma^2 = \frac{1}{C} \sum_{k=1}^C (x_k - \mu)^2 \quad (2)$$

Note that $\mu, \sigma \in \mathbb{R}$ and $y \in \mathbb{R}^C$. RMSNorm is very similar to LN but omits the centering,

$$y = \frac{x}{\sqrt{\tilde{\sigma}^2}} \quad (3)$$

where

$$\tilde{\sigma}^2 = \frac{1}{C} \sum_{k=1}^C x_k^2 = \frac{\|x\|^2}{C} \quad (4)$$

is the uncentered variance and $\|\cdot\|$ denotes the 2-norm. The impact of normalization on the training of neural networks as well as various alternative methods have been studied in many publications. A concise overview of the most important related work is given in App. A. Recently, [Zhu et al. \(2025\)](#) suggested an element-wise, non-linear transformation called Dynamic Tanh (DyT) as a drop-in replacement for LN:

$$y = \tanh(\alpha x) \quad (5)$$

It uses a learnable parameter $\alpha \in \mathbb{R}$, instead of relying on activation statistics like the traditional normalization methods discussed above. DyT is empirically well-motivated, as it was shown to resemble LN in the sense that it linearly transforms small values of x while squashing large values. However, the authors did not provide a theoretical justification for the observed similarity. The aim of our work is to fill this gap and to enhance the theoretical understanding of the relationship between LN and dynamic activation functions.

The paper is structured as follows. In Sec. 2, we show that DyT can be mathematically derived from RMSNorm using a well-defined decoupling in

derivative space. Sec. 3 discusses a similar element-wise transformation called *Dynamic Inverse Square Root Unit (DyISRU)*, which emerges from a decoupling procedure directly in function space. Sec. 4 demonstrates that DyISRU is more similar to normalization than DyT with regard to its effect on outliers. Our conclusions are presented in Sec. 5.

2 Dynamic Tanh (DyT)

In this section, we provide a mathematical derivation of the DyT function as an element-wise approximation of LN. Note that our starting point is actually RMSNorm, reflecting the fact that dynamic activation functions like Eq. (5) do not involve any subtractions for the purpose of centering. The derivation is done in three steps: (i) The derivative of the RMSNorm output with respect to its input is computed, resulting in a system of differential equations. (ii) The system of differential equations is decoupled and thereby reduced to a single equation. (iii) The decoupled differential equation is solved, leading to the DyT function. The process is illustrated in Fig. 1.

Theorem 1 (RMSNorm Derivative). *Let $x \in \mathbb{R}^C$ and*

$$y = \frac{x}{\sqrt{\tilde{\sigma}^2}} \quad (3)$$

with the uncentered variance of x ,

$$\tilde{\sigma}^2 = \frac{1}{C} \sum_{k=1}^C x_k^2 = \frac{\|x\|^2}{C} \quad (4)$$

Then $\forall i, j \in [1, \dots, C]$:

$$\frac{\partial y_i}{\partial x_j} = \frac{\sqrt{C}}{\|x\|} \left(\delta_{ij} - \frac{y_i y_j}{C} \right) \quad (6)$$

While this result has been obtained before (Xiong et al., 2020; Takase et al., 2025), we provide a proof using our notation in App. B.1.

Decoupling The partial derivatives of y_i in Eq. (6) depend on the different components of x in two ways. Firstly, through the off-diagonal elements ($i \neq j$) of the Jacobian. Secondly, all components x_k enter the expression for $\|x\|$. We can remove these cross-dependencies and decouple the system of differential equations by ignoring the off-diagonal terms and replacing the norm of x by a constant:

$$y_i y_j \rightarrow 0 \quad \forall i \neq j \quad (7)$$

$$\|x\| \rightarrow \frac{1}{\sqrt{\alpha}} \quad (8)$$

With these replacements, Eq. (6) becomes

$$\frac{\partial y_i}{\partial x_j} = \begin{cases} \sqrt{\alpha C} \left(1 - \frac{y_i^2}{C} \right) & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \quad (9)$$

We write the single differential equation on the diagonal succinctly as

$$\frac{dy_i}{dx_i} = \sqrt{\alpha C} \left(1 - \frac{y_i^2}{C} \right) \quad (10)$$

The following theorem shows that the DyT function can be obtained by solving Eq. (10).

Theorem 2 (DyT). *The differential equation*

$$\frac{dy_i}{dx_i} = \sqrt{\alpha C} \left(1 - \frac{y_i^2}{C} \right) \quad (10)$$

together with the initial condition

$$y_i(x_i = 0) = 0 \quad (11)$$

is solved by the function

$$y_i = \sqrt{C} \cdot \tanh(\alpha x_i) \quad (12)$$

The proof can be found in App. B.2. Eq. (12) represents the (scaled) DyT function. In contrast to the original formulation of DyT, Eq. (5), it explicitly contains the minimum and maximum value of y_i in terms of the scaling factor $\sqrt{C}$.

3 Dynamic Inverse Square Root Unit (DyISRU)

The results of the previous section raise the question whether it is possible to find an element-wise transformation akin to DyT that avoids both the detour through derivative space and neglecting off-diagonal elements of the Jacobian, Eq. (7). Such a function could be a more direct and accurate element-wise replacement for RMSNorm. To answer this question, we first rearrange the RMSNorm transformation, Eq. (3), and then decouple the components as before, see Fig. 1.

Theorem 3 (RMSNorm Rearrangement). *Let $x \in \mathbb{R}^C$ and*

$$y_i := \frac{x_i}{\sqrt{\tilde{\sigma}^2}} \quad (3)$$

be the output of the RMSNorm with

$$\tilde{\sigma}^2 = \frac{1}{C} \sum_{k=1}^C x_k^2 = \frac{1}{C} \|x\|^2 \quad (4)$$

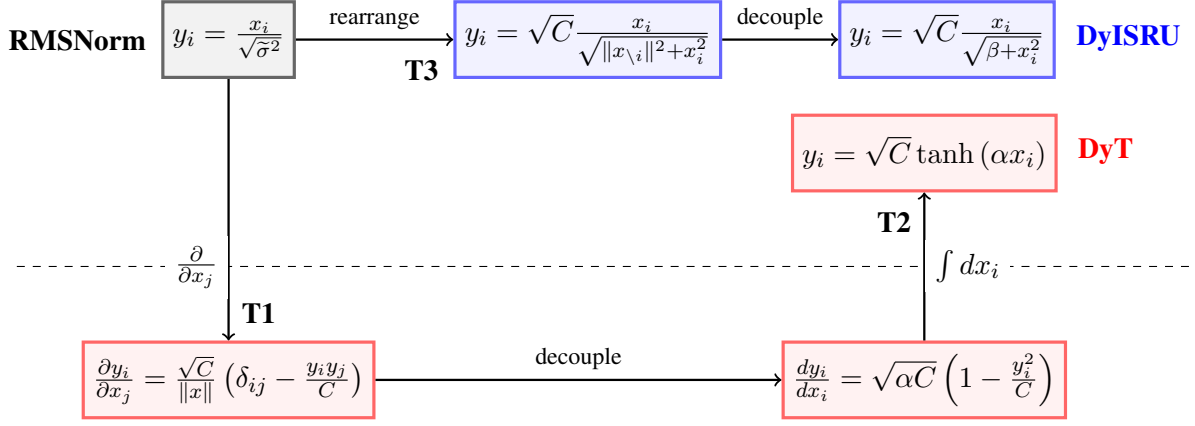


Figure 1: Illustration of how to obtain the dynamic activation functions DyT (red) and DyISRU (blue) from RMSNorm (black). The labels T1, T2, T3 indicate the application of our theorems. The dashed line differentiates between function space (y_i) above and derivative space ($\frac{\partial y_i}{\partial x_j}$) below.

being the uncentered variance of x . Eq. (3) can be written as

$$y_i = \sqrt{C} \cdot \frac{x_i}{\sqrt{\|x_{\setminus i}\|^2 + x_i^2}} \quad (13)$$

where

$$\|x_{\setminus i}\| := \sqrt{\sum_{k \neq i} x_k^2} \quad (14)$$

is the 2-norm of the vector $x_{\setminus i} \in \mathbb{R}^{C-1}$ comprised of all channels but i .

The proof can be found in App. B.3.

Decoupling The rearranged formulation of RMSNorm in Eq. (13) is convenient as the dependency on the components x_j with $j \neq i$ is isolated by means of $\|x_{\setminus i}\|$. An element-wise transformation can thus easily be obtained using the replacement

$$\|x_{\setminus i}\| \rightarrow \beta \quad (15)$$

where β is a global constant. It reads

$$y_i = \sqrt{C} \cdot \frac{x_i}{\sqrt{\beta + x_i^2}} \quad (16)$$

We call this function Dynamic Inverse Square Root Unit (DyISRU), as it is proportional to the ISRU activation function (Carlile et al., 2017), see App. C for details. Like the (scaled) DyT function in Eq. (12), it explicitly contains the minimum and maximum values of y_i in terms of $\sqrt{C}$. In App. D, we provide some considerations on how to use it in practice.

Comparison The two discussed element-wise transformations, DyT from Eq. (12) and DyISRU from Eq. (16), are compared in Fig. 2. Their shapes are quite similar, but DyT converges faster to the extrema ($\pm\sqrt{C}$) than DyISRU. Both dynamic activation functions employ learnable parameters instead of relying on activation statistics. However, DyT’s α globally models the inverse uncentered variance of x including all channels, Eq. (8), while DyISRU’s β globally models the uncentered variance of x including all channels but the transformed one, Eq. (15). More importantly, unlike DyISRU, DyT can be considered an approximation as it ignores the off-diagonal entries of the Jacobian, Eq. (7). This suggests that DyISRU more closely resembles normalization in comparison with DyT. In the next section, we will see that this is indeed the case.

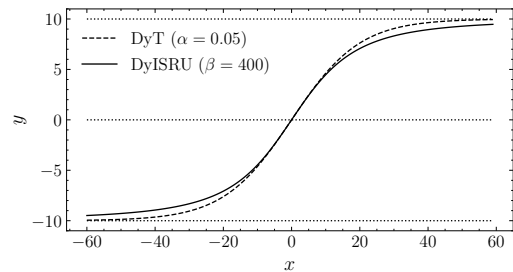


Figure 2: Functions DyT from Eq. (12) and DyISRU from Eq. (16) with parameters $\alpha = 0.05$ and $\beta = 400$ such that the derivatives at $x = 0$, namely α and $1/\sqrt{\beta}$, match. The dotted lines correspond to the extrema $y = \pm\sqrt{C}$.

4 Outlier Simulation

In order to compare RMSNorm and the considered dynamic activation functions, we first simulate data with outliers of different degrees and apply normalization. Afterwards, we employ DyT and DyISRU with optimal parameters α and β to investigate how well they describe the data.

4.1 Data: RMSNorm

We assume C channels and take a normally distributed sample of C values $x = (x_1, x_2, \dots, x_C)$ with zero mean and standard deviation σ :

$$x \sim \mathcal{N}(0, \sigma^2) \quad (17)$$

The parameters are set to arbitrary values, $C = 100$ and $\sigma = 2$. First, we apply RMSNorm and compute y according to Eq. (3). Next, we simulate outliers by iteratively increasing the largest value of x in steps of 5:

$$x_o \rightarrow x_o + 5 \cdot S \quad \text{with} \quad o = \underset{k}{\operatorname{argmax}} x_k \quad (18)$$

where S denotes the number of steps. At each step, RMSNorm is applied to the vector x to yield the output y . Repeating the process up to $S = 9$ yields the top plot of Fig. 3. In accordance with [Zhu et al. \(2025\)](#), we observe that (i) the slope of the linear function $y_i(x_i)$ decreases with the variance of x and (ii) the outliers follow a non-linear function. The larger the outlier x_o , the more squashed is the function.

4.2 Fit: DyT and DyISRU

In the next step, we aim to describe the simulated data by DyT and DyISRU as defined in Eq. (12) and (16), respectively. We only use the outliers (filled circles in Fig. 3) as data points for the fit¹, since it is primarily those we want to reproduce the normalization behavior for. Performing fits yields the optimal parameters

$$\alpha = 0.049 \quad \text{and} \quad \beta = 301.1 \quad (19)$$

for DyT and DyISRU, respectively. Both fitted functions are displayed in Fig. 3 together with the data and the residuals. We find that DyISRU describes the normalization data more accurately than DyT does. The mean absolute residuals are 0.33 for DyT and < 0.01 for DyISRU. This is consistent

¹In practice, we use the mirrored data points $(-x, -y)$ as well for the sake of numerical stability.

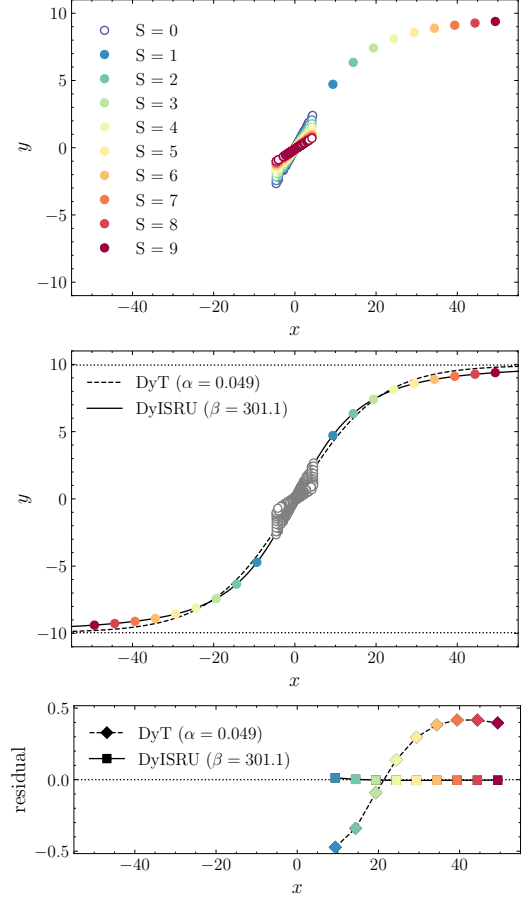


Figure 3: *Top:* Stepwise outlier simulation. The sample x and is plotted against its normalized counterpart y , with outliers of different degrees (filled circles) as defined by Eq. (18). *Center:* Functions DyT and DyISRU with optimal parameters α and β , respectively, fitted on the outliers shown as colored, filled circles. The non-outlier data are shown as gray, empty circles. *Bottom:* Residuals of the functions DyT and DyISRU with respect to the outlier data. As the residuals are antisymmetric (like the data and the functions), only positive outliers are displayed for the sake of simplicity.

with our theoretical finding that DyT and DyISRU correspond to approximate and exact element-wise counterparts of RMSNorm, respectively. Finally, we note that our results generalize to other distributions than the one used above, Eq. (17). Further details are provided in App. E.

5 Conclusions

This work provides a theoretical foundation for the empirically observed similarity of DyT and LN. We have detailed how dynamic activation functions emerge mathematically from RMSNorm by promoting channel-specific terms to a global, learnable parameter designed to describe outliers (decoupling). In addition, our analysis reveals that the exact element-wise counterpart to RMSNorm is given by a transformation called Dynamic Inverse Square

Root Unit (DyISRU), while DyT can be considered an approximation. The code used to reproduce our results is available at <https://github.com/flxst/dynamic-activation-functions>.

6 Limitations

While our work advances the theoretical understanding of dynamic activation functions, we did not conduct any experiments. Hence, we cannot make any statements about the practical implications of our work. In particular, it remains to be seen whether there is a notable difference between DyT and DyISRU in terms of model performance and with regard to the sensitivity on the learnable parameters’ initial values, which has been reported especially for LLMs (Zhu et al., 2025).

References

- Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. [Layer normalization](#).
- Andrew Brock, Soham De, and Samuel L. Smith. 2021a. [Characterizing signal propagation to close the performance gap in unnormalized resnets](#).
- Andy Brock, Soham De, Samuel L. Smith, and Karen Simonyan. 2021b. [High-performance large-scale image recognition without normalization](#). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1059–1071. PMLR.
- Brad Carlile, Guy Delamarter, Paul Kinney, Akiko Marti, and Brian Whitney. 2017. [Improving deep learning by inverse square root linear units \(isrlus\)](#).
- Soham De and Sam Smith. 2020. [Batch normalization biases residual blocks towards the identity function in deep networks](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 19964–19975. Curran Associates, Inc.
- Charles Dugas, Yoshua Bengio, François Bélisle, Claude Nadeau, and René Garcia. 2000. Incorporating second-order functional knowledge for better option pricing. *Advances in neural information processing systems*, 13.
- Bobby He and Thomas Hofmann. 2024. [Simplifying transformer blocks](#).
- Lei Huang, Jie Qin, Yi Zhou, Fan Zhu, Li Liu, and Ling Shao. 2023. [Normalization techniques in training dnns: Methodology, analysis and application](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10173–10196.
- Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. 2017. [Self-normalizing neural networks](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Kaifeng Lyu, Zhiyuan Li, and Sanjeev Arora. 2023. [Understanding the generalization benefit of normalization layers: Sharpness reduction](#).
- Yunhao Ni, Yuxin Guo, Junlong Jia, and Lei Huang. 2024. [On the nonlinearity of layer normalization](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 37957–37998. PMLR.
- Sho Takase, Shun Kiyono, Sosuke Kobayashi, and Jun Suzuki. 2025. [Spike no more: Stabilizing the pre-training of large language models](#).
- Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. 2020. [On layer normalization in the transformer architecture](#). In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10524–10533. PMLR.
- Biao Zhang and Rico Sennrich. 2019. [Root mean square layer normalization](#). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Hongyi Zhang, Yann N. Dauphin, and Tengyu Ma. 2019. [Fixup initialization: Residual learning without normalization](#).
- Jiachen Zhu, Xinlei Chen, Kaiming He, Yann LeCun, and Zhuang Liu. 2025. Transformers without normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

A Related Work

Our work is closely related to Zhu et al. (2025) and the idea of using dynamic activation functions to replace layer normalization in neural networks. However, other alternatives to normalization have been suggested in the literature. The arguably closest relatives to dynamic activation functions are self-normalizing neural networks (Klambauer et al., 2017), which use the special SELU activation function to enforce implicit normalization. Other notable work seeks to compensate for the absence of normalization by means of special initialization schemes (Zhang et al., 2019; De and Smith, 2020) and additional techniques (Brock et al., 2021a,b). Furthermore, approaches specific to transformers have been proposed, like alternative transformer block architectures that dispense with normalization (He and Hofmann, 2024). While we provide

insights into how layer normalization relates specifically to dynamic activation functions, the impact of layer normalization has been analyzed with regard to many different aspects, including but not limited to non-linearity (Ni et al., 2024) and generalization (Lyu et al., 2023).

B Theorem Proofs

B.1 Theorem 1

Theorem 1 (RMSNorm Derivative). *Let $x \in \mathbb{R}^C$ and*

$$y = \frac{x}{\sqrt{\tilde{\sigma}^2}} \quad (3)$$

with the uncentered variance of x ,

$$\tilde{\sigma}^2 = \frac{1}{C} \sum_{k=1}^C x_k^2 = \frac{\|x\|^2}{C} \quad (4)$$

Then $\forall i, j \in [1, \dots, C]$:

$$\frac{\partial y_i}{\partial x_j} = \frac{\sqrt{C}}{\|x\|} \left(\delta_{ij} - \frac{y_i y_j}{C} \right) \quad (6)$$

Proof. We start from Eq. (3) and compute the derivative of y_i with respect to x_j :

$$\frac{\partial y_i}{\partial x_j} = \frac{\partial}{\partial x_j} \left(\frac{x_i}{\sqrt{\tilde{\sigma}^2}} \right) \quad (20)$$

Defining

$$f = x_i \quad g = \sqrt{\tilde{\sigma}^2} \quad (21)$$

and using the shorthand notation

$$f' := \frac{\partial f}{\partial x_j} \quad g' := \frac{\partial g}{\partial x_j} \quad (22)$$

the quotient rule states

$$\frac{\partial y_i}{\partial x_j} = \frac{f'g - fg'}{g^2} \quad (23)$$

We compute the derivatives in Eq. (23):

$$\begin{aligned} f' &\stackrel{(21)}{=} \frac{\partial}{\partial x_j} x_i \\ &= \delta_{ij} \end{aligned} \quad (24)$$

and

$$\begin{aligned} g' &\stackrel{(21)}{=} \frac{\partial}{\partial x_j} \sqrt{\tilde{\sigma}^2} \\ &= \frac{1}{2\sqrt{\tilde{\sigma}^2}} \cdot \frac{\partial}{\partial x_j} (\tilde{\sigma}^2) \\ &\stackrel{(4)}{=} \frac{1}{2\sqrt{\tilde{\sigma}^2}} \cdot \frac{\partial}{\partial x_j} \left(\frac{1}{C} \sum_{k=1}^C x_k^2 \right) \\ &= \frac{1}{2C\sqrt{\tilde{\sigma}^2}} \cdot \frac{\partial}{\partial x_j} \left(x_j^2 + \sum_{k \neq j}^C x_k^2 \right) \\ &= \frac{1}{C\sqrt{\tilde{\sigma}^2}} \cdot x_j \\ &\stackrel{(3)}{=} \frac{y_j}{C} \end{aligned} \quad (25)$$

Inserting Eqs. (24) and (25) into Eq. (23), we get

$$\begin{aligned} \frac{\partial y_i}{\partial x_j} &\stackrel{(23)}{=} \frac{f'g - fg'}{g^2} \\ &\stackrel{(24,25)}{=} \frac{\delta_{ij} \cdot \sqrt{\tilde{\sigma}^2} - x_i \cdot \frac{y_j}{C}}{\tilde{\sigma}^2} \\ &\stackrel{(3)}{=} \frac{\delta_{ij} \cdot \sqrt{\tilde{\sigma}^2} - \frac{1}{C} \sqrt{\tilde{\sigma}^2} \cdot y_i y_j}{\tilde{\sigma}^2} \\ &= \frac{1}{\sqrt{\tilde{\sigma}^2}} \left(\delta_{ij} - \frac{y_i y_j}{C} \right) \end{aligned} \quad (26)$$

$$\stackrel{(4)}{=} \frac{\sqrt{C}}{\|x\|} \left(\delta_{ij} - \frac{y_i y_j}{C} \right) \quad (27)$$

□

B.2 Theorem 2

Theorem 2 (DyT). *The differential equation*

$$\frac{dy_i}{dx_i} = \sqrt{\alpha C} \left(1 - \frac{y_i^2}{C} \right) \quad (10)$$

together with the initial condition

$$y_i(x_i = 0) = 0 \quad (11)$$

is solved by the function

$$y_i = \sqrt{C} \cdot \tanh(\alpha x_i) \quad (12)$$

Proof. For the sake of readability, we temporarily drop the channel index i , i.e. we use $x_i \rightarrow x$ and $y_i \rightarrow y$. An additional slight rearrangement leads to

$$\frac{dy}{dx} = \frac{\sqrt{\alpha}}{\sqrt{C}} (C - y^2) \quad (28)$$

We separate the variables:

$$\begin{aligned}
\frac{\sqrt{\alpha}}{\sqrt{C}} \cdot dx &= \frac{dy}{C - y^2} \\
&= \frac{dy}{(\sqrt{C} - y)(\sqrt{C} + y)} \\
&= \frac{1}{2\sqrt{C}} \frac{\sqrt{C} - y + \sqrt{C} + y}{(\sqrt{C} - y)(\sqrt{C} + y)} dy \\
&= \frac{1}{2\sqrt{C}} \left(\frac{1}{\sqrt{C} + y} + \frac{1}{\sqrt{C} - y} \right) dy
\end{aligned} \tag{29}$$

Integration yields

$$\begin{aligned}
\frac{1}{2\sqrt{C}} \log \left(\frac{\sqrt{C} + y}{\sqrt{C} - y} \right) &= \frac{\alpha x}{\sqrt{C}} + \frac{c}{2\sqrt{C}} \\
\frac{\sqrt{C} + y}{\sqrt{C} - y} &= \exp(2\alpha x + c)
\end{aligned} \tag{30}$$

where c is an integration constant. Defining $Q := \exp(2\alpha x + c)$, we get

$$\begin{aligned}
\frac{\sqrt{C} + y}{\sqrt{C} - y} &= Q \\
\sqrt{C} + y &= (\sqrt{C} - y) Q \\
\sqrt{C} + y &= \sqrt{C} Q - Q y \\
Q y + y &= \sqrt{C} (Q - 1) \\
(Q + 1)y &= \sqrt{C} (Q - 1) \\
y &= \sqrt{C} \cdot \frac{Q - 1}{Q + 1}
\end{aligned} \tag{31}$$

Replacing Q again, and using $A = \exp(c)$, yields

$$y_i = \sqrt{C} \cdot \frac{A \exp(2\alpha x_i) - 1}{A \exp(2\alpha x_i) + 1} \tag{32}$$

Note that in the last equation, we have reintroduced the channel index i . We enforce the initial condition from Eq. (11) which requires $A = 1$. This leads to the scaled DyT function, Eq. (12). $\square$

B.3 Theorem 3

Theorem 3 (RMSNorm Rearrangement). *Let $x \in \mathbb{R}^C$ and*

$$y_i := \frac{x_i}{\sqrt{\tilde{\sigma}^2}} \tag{3}$$

be the output of the RMSNorm with

$$\tilde{\sigma}^2 = \frac{1}{C} \sum_{k=1}^C x_k^2 = \frac{1}{C} \|x\|^2 \tag{4}$$

being the uncentered variance of x . Eq. (3) can be written as

$$y_i = \sqrt{C} \cdot \frac{x_i}{\sqrt{\|x_{\setminus i}\|^2 + x_i^2}} \tag{13}$$

where

$$\|x_{\setminus i}\| := \sqrt{\sum_{k \neq i} x_k^2} \tag{14}$$

is the 2-norm of the vector $x_{\setminus i} \in \mathbb{R}^{C-1}$ comprised of all channels but i .

Proof.

$$\begin{aligned}
y_i &\stackrel{(3)}{=} \frac{x_i}{\sqrt{\tilde{\sigma}^2}} \\
&\stackrel{(4)}{=} \frac{x_i}{\sqrt{\frac{1}{C} \sum_{k=1}^C x_k^2}} \\
&= \sqrt{C} \cdot \frac{x_i}{\sqrt{\sum_{k \neq i} x_k^2 + x_i^2}} \\
&\stackrel{(14)}{=} \sqrt{C} \cdot \frac{x_i}{\sqrt{\|x_{\setminus i}\|^2 + x_i^2}}
\end{aligned} \tag{13}$$

$\square$

C Inverse Square Root Unit (ISRU)

The Inverse Square Root Unit (ISRU) function is defined in [Carlile et al. \(2017\)](#) as²

$$f(x) = \frac{x}{\sqrt{1 + \alpha x^2}} \tag{33}$$

This can also be written as

$$\begin{aligned}
f(x) &= \frac{x}{\sqrt{\alpha \cdot \left(\frac{1}{\alpha} + x^2\right)}} \\
&= \frac{\sqrt{\beta} x}{\sqrt{(\beta + x^2)}}
\end{aligned} \tag{34}$$

where we have identified

$$\beta := \frac{1}{\alpha} \tag{35}$$

in the second step. Eq. (34) is the same as Dy-ISRU from Eq. (16) apart from the factor $\sqrt{\beta}$ in the nominator.

²Note that the parameter α is not to be confused with the one used in DyT, Eq. (12).

D DyISRU in Practice

D.1 Initialization of β

Fig. 2 illustrates the similarity of DyT and DyISRU if their derivatives at the origin $x_i = 0$, given by

$$\left. \frac{dy_i}{dx_i} \right|_{x_i=0} \stackrel{(12)}{=} \alpha \quad (36)$$

for DyT and

$$\left. \frac{dy_i}{dx_i} \right|_{x_i=0} \stackrel{(16)}{=} \frac{1}{\sqrt{\beta}} \quad (37)$$

for DyISRU, are the same. A good starting point for the initialization β_0 can thus be inferred from values of α_0 that have been shown to work well in practice (Zhu et al., 2025), by matching Eq. (36) and (37). For example, the commonly used value $\alpha_0 = 0.5$ for DyT corresponds to $\beta_0 = \frac{1}{\alpha_0^2} = 4$ for DyISRU.

D.2 Parametrization of β

A potential practical problem of the DyISRU function from Eq. (16) is that it is only defined for $\beta > 0$. This can theoretically cause the training to fail, especially if the initial value β_0 is chosen to be small. A way to avoid instabilities is to parametrize β in terms of a differentiable, positive function $f : \mathbb{R} \rightarrow \mathbb{R}^+$, $\hat{\beta} \mapsto \beta$, like the softplus function (Dugas et al., 2000):

$$\beta = \text{softplus}(\hat{\beta}) = \log(1 + \exp(\hat{\beta})) \quad (38)$$

D.3 Efficiency

DyT has been shown to provide significant computational speed-ups compared to RMSNorm (Zhu et al., 2025). Since DyISRU is very similar to DyT in that it only uses element-wise transformations, we expect it to be about as fast. However, this question will eventually have to be answered by experiments.

E Outlier Simulation: Generalization

We extend the outlier simulation presented in Sec. 4 to other distributions, namely uniform and Laplace, with $\sigma \in \{1, 2, 4\}$. Tab. 1 shows that the results generalize well.

σ	distribution	MAR (DyT)	MAR (DyISRU)	α	β
1	normal	0.276	0.002	0.096	75.1
1	uniform	0.282	0.003	0.095	77.9
1	Laplace	0.275	0.002	0.096	74.7
2	normal	0.328	0.005	0.049	301.1
2	uniform	0.327	0.007	0.048	312.3
2	Laplace	0.327	0.005	0.049	299.4
4	normal	0.219	0.007	0.025	1206.6
4	uniform	0.203	0.010	0.025	1252.6
4	Laplace	0.221	0.007	0.026	1199.9

Table 1: Outlier simulation results for different distributions. MAR stands for mean absolute residual, see Sec. 4.